High prevalence of cancer-associated TP53 variants in the gnomAD database: A word of caution concerning the use of variant filtering

Thierry Soussi, Bernard Leroy, Michal Devir, Shai Rosenberg

Human Mutation

High prevalence of cancer-associated TP53 variants in the gnomAD database: a word of caution concerning the use of variant filtering.

Journal:	*Human Mutation*
Manuscript ID	humu-2018-0526.R1
Wiley - Manuscript type:	Databases
Date Submitted by the Author:	n/a
Complete List of Authors:	Soussi, Thierry; Karolinska Institutet, Dept. of Oncology-Pathology Bernard, Leroy; Université Pierre et Marie Curie-Paris 6, Devir, Michal; Hebrew University Medical Center Jerusalem, Laboratory for Cancer Computational Biology, Hadassah Rosenberg, Shai; Hebrew University Medical Center Jerusalem, Laboratory for Cancer Computational Biology, Hadassah
Key Words:	TP53, SNP database, gnomAD, pathogenicity

High prevalence of cancer-associated TP53 variants in the gnomAD database: a word of caution concerning the use of variant filtering.

Thierry Soussi [1,2,3], Bernard Leroy[1], Michal Devir [4] and Shai Rosenberg [4,5]

[1] Sorbonne Université, UPMC Univ Paris 06, F- 75005 Paris, France

[2] INSERM, U1138, Centre de Recherche des Cordeliers, Paris, France

[3] Department of Oncology-Pathology, Cancer Center Karolinska (CCK), Karolinska Institutet, Stockholm, Sweden

[4] Laboratory for Cancer Computational Biology, Hadassah – Hebrew University Medical Center Jerusalem, 91120, Israel. POB 12000.

[5] Center for Neurooncology Hadassah – Hebrew University Medical Center. Jerusalem, 91120, Israel. POB 12000

Correspondence: thierry.soussi@ki.se

Keywords: TP53 variants; SNP; gnomAD: ExAc: variant pathogenicity

Grant Sponsor: Radiumhemmets Forskningsfonder to TS.

Abstract

The 1000 genome project, the Exome Aggregation Consortium (ExAC) or the Genome Aggregation database (gnomAD) datasets, were developed to provide large-scale reference data of genetic variations for various populations to filter out common benign variants and identify rare variants of clinical importance based on their frequency in the human population.

Using a TP53 repository of 80,000 cancer variants, as well as *TP53* variants from multiple cancer genome projects, we have defined a set of certified oncogenic *TP53* variants. This specific set has been independently validated by functional and *in silico* predictive analysis.

Here we show that a significant number of these variants are included in GnomAD and ExAC. Most of them correspond to TP53 hot spot variants occurring as somatic and germline events in human cancer.

Similarly, disease-associated variants for 5 other tumor suppressor genes, including *BRCA1, BRCA2, APC, PTEN* and *MLH1*, have also been identified.

This study demonstrates that germline *TP53* variants in the human population are more frequent than previously thought. Furthermore, population databases such as gnomAD or ExAC must be used with caution and need to be annotated for the presence of oncogenic variants in order to improve their clinical utility.

Introduction

Predicting the pathogenicity of genetic variants associated with various types of diseases constitutes a major challenge for precision medicine (Evans et al., 2017, Niroula and Vihinen, 2017). Assessment of constitutional variants of genes associated with hereditary diseases requires multiple lines of evidence, including the demonstration that the variant is not a non-pathogenic variant present in the population. The situation is less complex in the context of cancer, which is usually associated with somatic variants, as tumor-specific variants can be identified by comparison with matched non-tumor DNA from the patients, when available. However, as tumors are associated with many somatic variations, it is crucial to distinguish true driver variants from the majority of passenger-neutral variants. Multiple databases describing cancer genes are now available and include data from tens of thousands of cancer patients (Yang et al., 2015). Although constitutional variants of many cancer genes (mostly oncogenes) have rarely been identified, germline variants in several tumor suppressor genes, such as *APC* (APC (MIM: 191170) or *PTEN* (MIM: 601728), are associated with familial cancer. The pathogenicity of variants of these genes occurring as germline and somatic events can now be more easily predicted on the basis of data derived from various cancer variants databases.

Regardless of the type of disease, the first step in predicting the deleteriousness of a genetic variant requires evaluation of the frequency of the variant in the general population using various human genetic population databases.

Single-nucleotide polymorphisms (SNPs) represent the most common form of genetic variation in the human genome. SNPs were originally defined as constitutional variations present at a frequency of at least 1% in the general population (http://www.ncbi.nlm.nih.gov/books/NBK21088/) (Brookes, 1999, Tennessen et al., 2012). SNPs are the mainstay of diversity in species and have been found to be tremendously useful

markers for genetic studies. The dbSNP database, created in 1998 and maintained by the NIH, keeps track of SNPs (http://www.ncbi.nlm.nih.gov/SNP/)(Sherry et al., 2001).

The advent of massively parallel sequencing (next-generation sequencing, or NGS) has ushered in a new era of analysis of the human population with the recognition that the human genome includes far more SNP than initially suspected, including many non-pathogenic variants at frequencies well below the 1% threshold previously used (Karki et al., 2015). Indeed, the last build of dbSNP (build 151) includes 325 million variants, but many studies have questioned the quality of the various entries (Arthur et al., 2015, Musumeci et al., 2010). In contrast, data from the 2,504 individuals of the 1000 genome project have been highly curated and are frequently used as a reference to remove common genuine variants, as recommended by the American College of Medical Genetics and Genomics (1000 et al., 2015, Richards et al., 2015). Nevertheless, the small number of entries as well as the high geographical diversity of this cohort (26 populations, 5 super-populations) do not allow full coverage of low frequency SNPs (http://www.internationalgenome.org/data/).

In 2014, the Exome Aggregation Consortium (ExAC) released exome data from 60,706 individuals (Lek et al., 2016). This new database was designed to be used as a novel reference set for SNP allele frequency. ExAC includes data from the 1,000 genomes project, as well as constitutional data from unrelated individuals sequenced as part of various disease-specific and population genetic studies. In 2016, ExAC was transformed into The Genome Aggregation Database (gnomAD) with the inclusion of new data, leading to a dataset with information from 123,136 exome sequences and 15,496 whole-genome sequences of unrelated individuals (http://gnomad.broadinstitute.org/). ExAC and gnomAD have both been widely used as a substitute or complement of dbSNP and are currently used in multiple analytical pipelines.

One disadvantage of sequencing large population cohorts is the incidental identification of a rare pathogenic germline variants unrelated to the goal of the study (Blackburn et al., 2015). Inclusion of these variants could lead to misinterpretation and these variants should therefore be annotated.

In the present study, we analyzed gnomAD for the presence of pathogenic *TP53* variants, based on the availability of multiple, well-defined *TP53* mutation databases (Leroy et al., 2017). The first step, using an independent set of cancer-associated variants from various origins, consisted of validating a certified set of pathogenic *TP53* variants. The gnomAD was then examined to identify multiple pathogenic *TP53* variants, including several hot spot variants found in many cancers. Finally, five other tumor suppressor genes were also examined and pathogenic germline variants were identified in the genomAD.

Material and Methods

Database analysis

The current version of the UMD_TP53 database includes 80,406 variants identified in tumors, cell lines (somatic variants) or in patients with hereditary cancers (germline variants) (database freeze Oct 2017) (Leroy et al., 2017). As most variants are described multiple times in different patients, the database includes 6,872 unique TP53 variants. This database includes *TP53* alterations identified by conventional sequencing, as well as alterations identified in next-generation sequencing-based projects such as TCGA (https://p53.fr). A subset derived from this database includes only those studies using conventional Sanger sequencing for the detection of *TP53* variants. This dataset

comprises 37,295 variants (4,299 unique variants). The use of a specific Sanger dataset avoids redundancies and ensures the specificity of the three datasets described below.

TP53 variants from various databases were downloaded from their respective portals **(Supplementary Table S1)**. TCGA and MSKCC TP53 variants were downloaded from the cbioportal (http://www.cbioportal.org/, October 2017); ICGC TP53 variants were downloaded from the ICGC data portal (https://dcc.icgc.org/, data release 26, Dec, 17th 2017).

GnomAD data (r2.02, Oct 2017) and ExAC data (0.31), were obtained via http://gnomad.broadinstitute.org and ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/subsets/, respectively. For the ExAC database, we used both the full dataset as well as the curated dataset that does not include TCGA data. We observed a number of discrepancies in the coordinates used to define the boundaries of the *TP53* gene **(Supplementary Table S2)**. Although ExAC and gnomAD used the coordinates defined by Ensembl (ENSG00000141510, 7,661,779-7,687,550 GRCh38, with a *TP53* gene size of 25,771 bp), other databases use the official coordinates as defined by the Locus Reference Genomic (http://ftp.ebi.ac.uk/pub/databases/lrgex/LRG_321.xml) based on RefSeq gene NG_017013.2 in Genbank (7,692,550- 7,659,779 GRCh38 with a *TP53* gene size of 32,772 bp). To ensure accurate comparison between the various datasets, data from gnomAD and ExAC were obtained by using the genomic coordinates of the entire TP53 gene defined by RefSeq **(Supplementary Table S2)**.

Recent studies have shown that cancer gene variants can be found in mature blood cells of approximately 10% of individuals aged >65 years due to clonal hematopoiesis, (CH) a clonal expansion of a single mutant hematopoietic cell . Population aggregated in gnomAD as well as in ExAc are issued from multiple studies, some of them performed more than 10 years ago when the importance of CH was underappreciated. TP53 is one of the genes

frequently mutated in CH, which could lead to false results in SNP analysis, as DNA is often extracted from peripheral blood cells. Nevertheless, in most cases, the allele frequency of these variants in CH is low (less than 20%), less than the 50% observed for a normal allelic distribution of a SNP. Depending on the stringency of the criteria used to analyse and filter the SNP data, it is possible that some of these variant found in population database are indeed due to CH.

In the first step, minimal genomic data, such as genomic coordinates and genetic events, were extracted from each dataset to define the correct annotation according to HGVS recommendations. In a second step, variant annotation was validated by using the Name Checker tool developed by Mutalyzer (https://mutalyzer.nl/). Mutalyzer handles all types of variations targeting the TP53 gene, such as substitutions, insertions, duplications, deletions, or more complex insertion/deletion (Wildeman et al., 2008). The current version of Mutalyzer (Mutalyzer 2.0.26) uses the stable NCBI sequence NG_017013.2 as reference for TP53. This is a key issue in order to avoid problems associated with the use of multiple genome references (NCBI Build 36.1/hg18, Genome Reference Consortium GRCh37/hg19 or Genome Reference Consortium GRCh38/hg38) by the various NGS pipelines as well as non-compliant nomenclatures.

Variant effect prediction

dbNSFP v3.5 was downloaded from (https://sites.google.com/site/jpopgen/dbNSFP). It compiles prediction or conservation scores from multiple prediction algorithms (Liu et al., 2016). Prediction scores from each algorithm were available as normalized data from 0 (less deleterious) to 1 (most deleterious). Envision scores for all TP53 missense variants

were downloaded from https://envision.gs.washington.edu/shiny/envision_new/ (Gray et al., 2018). Envision algorithm is based on variant effect measurements from multiple large-scale mutagenesis datasets. Each variant was annotated with 27 features designed to describe evolutionary, structural or physicochemical characteristics of the residue. Envision scores were available as normalized data from 1 (less deleterious) to 0 (most deleterious).

The three datasets analyzed correspond to all *TP53* variants found in the Sanger dataset (Sanger), in CSD (CSD_IN) or in the Sanger dataset without CSD variants (CSD_OUT). Box and whisker plots show the upper and lower quartiles and range (box), median value (horizontal line inside the box), and 1-99 percentile (whisker line). For SIFT, CADD and REVEL, higher scores are associated with deleterious variants, whereas the opposite applies to ENVISION scores. D: Boxplot plot showing TP53 variant loss of activity in the various datasets.

The activity of TP53 protein variants has been described in detail in a previous report (Kato et al., 2003, Soussi et al., 2005). Briefly, haploid yeast transformants containing 2,314 TP53 variants and a green fluorescent protein reporter plasmid were constructed. TP53 activity was tested by measuring the fluorescent intensity of green fluorescent protein that is controlled by various promoter sequences regulated by TP53 after 3 days of growth at 37°C. Wild-type TP53 activity for each promoter was set at 100% and empty plasmid was set at 0.

Results

Defining a curated set of pathogenic TP53 variants.

Four different non-overlapping datasets of *TP53* variants were used to define a curated set of pathogenic *TP53* variants **(Figure 1 and supplementary Table S3)**. The Sanger dataset includes 4,299 variants extracted from the latest version of the UMD_TP53 mutation database (Leroy et al., 2014). All these variants were described in studies using Sanger sequencing and were published between 1989 and 2017. The other three sets of *TP53* variants are derived from large independent cancer sequencing projects: TCGA (Network et al., 2013)., ICGC (Hudson et al., 2010). and Zehir's study from the MSK-IMPACT project (Zehir et al., 2017). Each of these studies described more than 1,000 different TP53 variants at various frequencies in the coding regions or splice sites (+2/-2) of *TP53* **(Supplementary Table S3)**. A novel set of TP53 variants designated as "cancer shared dataset" (CSD) was created by combining the above four datasets to define a core of 471 recurrent TP53 variants found at least once in each database **(Figure 2A).** As these four datasets are derived from independent studies using different patients and different methodologies, it is highly likely that these 471 shared variants are true recurrent pathogenic variants. The frequency of each of these variants in the four datasets is strikingly similar and includes all the major *TP53* hotspot variants **(Supplementary Table S4).** The only difference concerns the higher frequency of variants in exons 10 and 11 for the NGS data, as these regions were rarely screened by Sanger sequencing.

The CSD includes 168 variants that lead to the expression of truncated proteins (59 nonsense variants, 53 splice variants, 49 indels and 7 in-frame indels), which are most likely deleterious for TP53 growth suppressor activity. The remaining 303 single nucleotide

variants include 298 missense variants and 5 synonymous variants. The five synonymous variants are localized in the last base of exons 4, 6 and 9 and are known to drastically impair *TP53* splicing **(Supplementary Table S4)** (Soussi et al., 2017, Supek et al., 2014)**..** The CSD was compared to the 1,144 TP53 variants from the NCBI ClinVar database **(Figure 2B).** ClinVar is a freely available, central database including genes and variants of clinical importance and corresponding clinical and experimental evidence for a wide range of genes and related disorders (Landrum et al., 2018). Two hundred eighty-six of the 471 TP53 variants found in the CSD are also included in ClinVar with 195 (68%) considered to be either pathogenic or likely pathogenic **(Figure 2B)**. The remaining variants have been labeled as either uncertain or conflicting. None of the 246 benign TP53 variants described in ClinVar were found in the cancer shared dataset, an observation that reinforces the quality of this dataset.

The pathogenicity of the missense variants included in the CSD has also been assessed using the prediction scores from dbNSFP (see Material and methods) (Liu et al., 2016). **(Figure 2C and Supplementary Figure S1)**. The rank scores of *TP53* variants from the cancer shared dataset are always associated with a high pathogenic score significantly different from the scores of the TP53 variants not included in this dataset **(Figure 2C and Supplementary Figure S1**). Envision, a novel predictor algorithm that has been shown to outperform conventional predictors, was also used and provided similar results (Gray et al., 2018). **(Figure 2C)**.

A unique feature of the TP53 mutation database compared to other mutation databases is the availability of quantitative functional data for all missense variants **(Figure 2D).** The combination of TP53 variant frequency and functional information in the database clearly highlights a marked inverse correlation between the frequency of TP53 mutants and their activity: frequent TP53 mutants are always inactive (activity less than 20% compared to

wild-type TP53), whereas approximately one-half of the mutants reported only once have an activity greater than 50% compared to wild-type TP53 **(Figure 2D)**. This result is in total agreement with our previous analysis of the 2014 release of the UMD_TP53 database (Leroy et al., 2014). *TP53* variants from the cancer shared dataset also display a significant loss of activity similar to that of the groups of more frequent *TP53* variants in the UMP_TP53 dataset (**Figure 2D**). Individual analysis of the residual activity of the 298 missense variants identified in the cancer shared dataset revealed that most of these variants are fully defective for 6 different *TP53* response elements compared to the other missense variants in the database **(Supplementary Figure S2A and B).**

Overall, the cancer shared dataset identified in this analysis includes a high quality set of frequent pathogenic TP53 variants that can be used as a reference for further analysis.

Oncogenic TP53 variant are included in GnomAD and ExAc.

A total of 3,301 constitutional *TP53* variants were extracted from gnomAD, including 307 variants in the coding region of the major transcript (NM_000546). These variants were first compared with data from ClinVar. Although many variants were absent, as expected, the remaining variants were classified as either Likely Benign or Uncertain, while only a few were classified as Pathogenic or Likely Pathogenic **(Supplementary Figure S3A and B**). Functional analysis of these variants showed that most of them are still active, confirming that they most correspond to benign variants **(Figure 2D)**. GnomAD variants were then compared to the CSD. Thirty-seven *TP53* variants were present in both datasets, including 34 missense variants, 2 nonsense variants, and 1 frameshift variant **(Figure 1)**. Despite the low frequency of each of these variants in the gnomAD database (ranging from 1 to 10 among the 37 variants leading to a total of 73 individuals with one of these variants), several of these variants correspond to hotspot variants described in

human cancer and known to be fully defective for TP53 tumor suppressor activity **(Figure 3A)**. *TP53* variants, such as NM_000546_c.524G>A; NP_000537_p.Arg175His, NM_000546_c.7817C>T; NP_000537_p.Arg273Cys or NM_000546_743G>A; NP_000537_p.Arg248Gln, have also been shown to be highly oncogenic both *in vitro* and *in vivo* in mice knock-in models with multiple gain of function (Jackson and Lozano, 2013, Oren and Rotter, 2010).

Heat map of TP53 activity shows that the majority of the 34 missense variants are indeed defective for TP53 transactivating activity **(Figure 3).** Variant NM_000546_c.672G>T; NP_000537_p.Glu224Asp does not display any loss of transcriptional activity, but this single nucleotide substitution, localized in the last base of exon 6, is known to be deleterious for *TP53* splicing, leading to a *TP53* null phenotype, as the putative mutant protein is not expressed (Supek et al., 2014). This observation illustrates the limitation of certain analytical tools, which predicted that an amino acid change affects protein function. However, it is also possible that a missense variant does not act at the protein level, but rather at the nucleotide level by interfering with the correct assembly of the pre-mRNA splicing machinery. The correct nomenclature for this variant should be NM_000546_c.672G>T; NP_000537_p.?, as recommended by HGVS.

Many different computational variant effect predictors all predicted these variants to be pathogenic despite using algorithms based on protein structure or sequence conservation (**Supplementary Figure S4**). None of these variants were predicted to be benign according to ClinVar **(Supplementary Figure S3C).**

The gnomAD database includes data from the ExAC database (an aggregation from 60,706 individual exome analyses) and was subsequently completed with data from 123,136 individual exomes. It also includes whole-genome sequencing data from 15,496 individuals. However, several patient cohorts from ExAc were not included in gnomAD

(http://gnomad.broadinstitute.org/). Analysis of *TP53* variants from ExAc shows that 28 of the 34 missense variants are present in both datasets at similar frequencies. New variants in gnomAD are likely to be derived from the new datasets that have been added. We also observed that 8 *TP53* variants identified solely in ExAc are included in the CSD **(Figure 3A)**. Whether these variants belong to ExAC-specific patient cohorts or have been filtered out in gnomAD is currently unknown.

The gnomAD and ExAC datasets have a highly heterogeneous origin, including multiple studies on various diseases. Notably, these datasets include germline data from patients used by the TCGA consortium for cancer genome sequencing. Due to the association between germline *TP53* variants and various cancer predisposition syndromes, it is likely that several pathogenic *TP53* variants observed in gnomAD could be derived from the TCGA dataset. Although TCGA data cannot be removed from gnomAD, a version of the ExAC database that does not include TCGA data (cancer-free) is available and was examined for TP53 variants **(Figure 3B)**. Ten pathogenic *TP53* variants were filtered out in ExAc when TCGA data were removed, confirming the high prevalence of *TP53* germline mutations in the TCGA dataset **(Figure 3B)** (Huang et al., 2018)**.** These variants are still included in gnomAD.

One hundred thirty-five (135) missense *TP53* variants found in gnomAD were not detected in the cancer shared dataset, although they are present at various frequencies in the Sanger dataset **(Supplementary Figure S5).** Whether these variants constitute low-frequency TP53 pathogenic variants or uncommon benign constitutional variants remains an open question.

Altogether, our analysis demonstrates that pathogenic *TP53* variants are included in both gnomAD and ExAC with a certain degree of heterogeneity between the two datasets.

Pathogenic variants for other tumor suppressor genes.

To determine whether our observation is specific for *TP53*, the presence of pathogenic variants for other cancer genes was also analyzed. Five tumor suppressor genes known to harbor both somatic and germline mutations in various cancer types were selected: *BRCA1* (MIM: 113705), *BRCA2* (MIM: 600185), *APC* (MIM611731), *MLH1* (mutL homolog 1 (MIM: 120436) and *SKT11* (MIM: 602216).

For each gene, gnomAD data were crossed with ClinVar data and frequency data from the COSMIC database **(Table 1 and Supplementary Table S5)**. Pathogenic or likely pathogenic variants of all genes, as defined by ClinVar, were identified in gnomAD. Several variants, such as NM_000314.6(PTEN):c.388C>T (p.Arg130Ter), NM_000038.5(APC):c.3927_3931delAAAGA (p.Glu1309Aspfs) or NM_007294.3(BRCA1):c.68_69delAG (p.Glu23Valfs), are known to be deleterious and are found as both somatic mutations in sporadic cancer and germline variants in hereditary syndromes.

Discussion

Stringent criteria were used in this study to identify potential pathogenic variants in the gnomAD database, which is the largest dataset of variants detected in the human population. We first identified and validated a set of pathogenic *TP53* variants using multiple independent cancer databases. When compared to gnomAD, this cancer shared dataset helped to identify pathogenic *TP53* variants in gnomAD, including 14 variants previously defined as pathogenic or likely pathogenic in ClinVar.

The results of this study highlight two different issues: firstly, the frequency of *TP53* pathogenic variants in the general population.

The frequency of pathogenic germline *TP53* variants in the general population has been estimated to be between 1 in 5,000 and 1 in 20,000 (Gonzalez et al., 2009b, Peng et al., 2017). At least 15 to 20% of these germline variants have been shown to be *de novo* mutations, making large-scale studies more complicated, as the lack of family history of cancer will lead to sampling bias for population analysis (Gonzalez et al., 2009a, Renaux-Petel et al., 2018).

By taking into account the incidence of each variant found in the various datasets, the frequency of germline *TP53* variants was 1/831, 1/1.235 and 1/1.899 in ExAC, ExAC without TCGA and gnomAD datasets, respectively. Although the high frequency of pathogenic TP53 variants in ExAC is biased by the TCGA data, the other two datasets should provide a more accurate picture of the frequency of TP53 germline mutations. In view of the highly stringent criteria that we have developed for the detection of pathogenic TP53 variants, it is likely that other less frequent pathogenic variants may also be included in gnomAD. Historically, germline TP53 variants have been identified in Li-Fraumeni syndrome (LFS), a very rare inherited familial predisposition to a wide range of often rare cancers (Malkin et al., 1990). It was subsequently observed that germline TP53 mutations can be associated with other familial cancers, such as breast/ovarian cancer or sarcoma, indicating that these germline TP53 variants are not restricted to this rare syndrome and are more frequent than previously thought (McCuaig et al., 2012). Despite the high frequency of germline TP53 variants observed in our analysis, the penetrance of these variants is likely to be highly heterogeneous according to the type of variant and the individual's genetic background and lifestyle. Large case-control studies are necessary to fully evaluate the cancer risks associated with all of these mutations.

In the present study, the variants included in all population databases were considered to be true germline alleles derived from both parents. Recent studies have shown that somatic TP53 variants can arise in the hematologic compartment, leading to clonal

hematopoiesis (CH), which is usually observed in older individuals. A high predominance of these clones and insufficiently rigorous SNP analysis can lead to false interpretation. Nevertheless, most TP53 variants found in CH have an allelic frequency less than 20%, which should avoid classification of these variants as germline mutations (Mitchell et al., 2018).

The second issue raised by our results concerns the presence in both the gnomAD and ExAC datasets of TP53 pathogenic variants, as well as pathogenic variants of other genes. In a previous work, Kobayashi et al. identified 14 pathogenic BRCA1 variants in the ExAc database (Kobayashi et al., 2017). This problem can introduce significant errors in genomic studies, as the ExAC and gnomAD databases are both used in pipelines to filter common or uncommon SNPs in human populations (Huang et al., 2018). Both datasets have also been used as a negative, non-pathogenic sets for the purposes of training predictive algorithms (Jagadeesh et al., 2016). **Figure 4** summarizes the analysis performed in this study and displays the population frequency of all missense variants found in gnomAD. Although most pathogenic variants are found at frequencies below 0,00005, the use of a threshold of 0.0001 should ensure that no pathogenic variant is selected.

Because these databases are invaluable tools for the detection of rare and/or population-specific SNPs, we believe that careful curation is mandatory in order to annotate these variants. As such curation cannot be performed on a large scale by any global predictive tools, we suggest that this procedure should be performed for each gene by a team of specialized curators, as illustrated in the present study.

This type of initiative will improve the quality of population datasets, as well as future databases that will emerge from the wealth of information provided by the multiple genome sequencing initiatives launched all around the world.

Supplemental Data

Online supplemental Data include 5 figures and 5 tables.

Funding

TS is supported by Radiumhemmets Forskningsfonder and the Swedish Cancer Society (Cancerfonden) .

References

1000 GPC, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. Nature 526: 68-74.

Arthur JW, Cheung FS, Reichardt JK. 2015. Single nucleotide differences (SNDs) continue to contaminate the dbSNP database with consequences for human genomics and health. Hum Mutat 36: 196-199.

Blackburn HL, Schroeder B, Turner C, Shriver CD, Ellsworth DL, Ellsworth RE. 2015. Management of Incidental Findings in the Era of Next-generation Sequencing. Curr Genomics 16: 159-174.

Brookes AJ. 1999. The essence of SNPs. Gene 234: 177-186.

Network CGAR, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45: 1113-1120.

Evans JP, Powell BC, Berg JS. 2017. Finding the Rare Pathogenic Variants in a Human Genome. JAMA 317: 1904-1905.

Gonzalez KD, Buzin CH, Noltner KA, Gu D, Li W, Malkin D, Sommer SS. 2009a. High frequency of de novo mutations in Li-Fraumeni syndrome. J Med Genet 46: 689-693.

Gonzalez KD, Noltner KA, Buzin CH, Gu D, Wen-Fong CY, Nguyen VQ, Han JH, Lowstuter K, Longmate J, Sommer SS, Weitzel JN. 2009b. Beyond Li Fraumeni Syndrome: clinical characteristics of families with p53 germline mutations. J Clin Oncol 27: 1250-1256.

Gray VE, Hause RJ, Luebeck J, Shendure J, Fowler DM. 2018. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. Cell Syst 6: 116-124.e3.

Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, Scott AD, Krassowski M et al. 2018. Pathogenic Germline Variants in 10,389 Adult Cancers. Cell 173: 355-370.e14.

Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M et al. 2010. International network of cancer genome projects. Nature 464: 993-998.

Jackson JG, Lozano G. 2013. The mutant p53 mouse as a pre-clinical model. Oncogene 32: 4325-4330.

Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein JA, Bejerano G. 2016. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet 48: 1581-1586.

Karki R, Pandya D, Elston RC, Ferlini C. 2015. Defining “mutation” and “polymorphism” in the era of personal genomics. BMC Med Genomics 8: 37.

Kato S, Han SY, Liu W, Otsuka K, Shibata H, Kanamaru R, Ishioka C. 2003. Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. Proc Natl Acad Sci U S A 100: 8424-8429.

Kobayashi Y, Yang S, Nykamp K, Garcia J, Lincoln SE, Topper SE. 2017. Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. Genome Med 9: 13.

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K et al. 2018. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46: D1062-D1067.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O’Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536: 285-291.

Leroy B, Anderson M, Soussi T. 2014. TP53 mutations in human cancer: database reassessment and prospects for the next decade. Hum Mutat 35: 672-688.

Leroy B, Ballinger ML, Baran-Marszak F, Bond GL, Braithwaite A, Concin N, Donehower LA, El-Deiry WS, Fenaux P, Gaidano G, Langerød A, Hellstrom-Lindberg E et al. 2017. Recommended Guidelines for Validation, Quality Control, and Reporting of TP53 Variants in Clinical Practice. Cancer Res 77: 1250-1260.

Liu X, Wu C, Li C, Boerwinkle E. 2016. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat 37: 235-241.

Malkin D, Li FP, Strong LC, Fraumeni JFJ, Nelson CE, Kim DH, Kassel J, Gryka MA, Bischoff FZ, Tainsky MA, Friend S. 1990. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science 250: 1233-1238.

McCuaig JM, Armel SR, Novokmet A, Ginsburg OM, Demsky R, Narod SA, Malkin D. 2012. Routine TP53 testing for breast cancer under age 30: ready for prime time? Fam Cancer 11: 607-613.

Mitchell RL, Kosche C, Burgess K, Wadhwa S, Buckingham L, Ghai R, Rotmensch J, Klapko O, Usha L. 2018. Misdiagnosis of Li-Fraumeni Syndrome in a Patient With Clonal Hematopoiesis and a Somatic TP53 Mutation. J Natl Compr Canc Netw 16: 461-466.

Musumeci L, Arthur JW, Cheung FS, Hoque A, Lippman S, Reichardt JK. 2010. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. Hum Mutat 31: 67-73.

Niroula A, Vihinen M. 2017. Predicting Severity of Disease-Causing Variants. Hum Mutat 38: 357-364.

Oren M, Rotter V. 2010. Mutant p53 gain-of-function in cancer. Cold Spring Harb Perspect

Biol 2: a001107.

Peng G, Bojadzieva J, Ballinger ML, Li J, Blackford AL, Mai PL, Savage SA, Thomas DM, Strong LC, Wang W. 2017. Estimating TP53 Mutation Carrier Probability in Families with Li-Fraumeni Syndrome Using LFSPRO. Cancer Epidemiol Biomarkers Prev 26: 837-844.

Renaux-Petel M, Charbonnier F, Théry JC, Fermey P, Lienard G, Bou J, Coutant S, Vezain M, Kasper E, Fourneaux S, Manase S, Blanluet M et al. 2018. Contribution of de novo and mosaic TP53 mutations to Li-Fraumeni syndrome. J Med Genet 55: 173-180.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17: 405-424.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-311.

Soussi T, Kato S, Levy PP, Ishioka C. 2005. Reassessment of the TP53 mutation database in human disease by data mining with a library of TP53 missense mutations. Hum Mutat 25: 6-17.

Soussi T, Taschner PE, Samuels Y. 2017. Synonymous Somatic Variants in Human Cancer Are Not Infamous: A Plea for Full Disclosure in Databases and Publications. Hum Mutat 38: 339-342.

Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. 2014. Synonymous mutations frequently act as driver mutations in human cancers. Cell 156: 1324-1335.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D et al. 2012. Evolution and functional impact of rare

coding variation from deep sequencing of human exomes. Science 337: 64-69.

Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. 2008. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 29: 6-13.

Yang Y, Dong X, Xie B, Ding N, Chen J, Li Y, Zhang Q, Qu H, Fang X. 2015. Databases and web tools for cancer genomics study. Genomics Proteomics Bioinformatics 13: 46-50.

Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, Srinivasan P, Gao J, Chakravarty D, Devlin SM, Hellmann MD, Barron DA et al. 2017. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med 23: 703-713.

Figure legends

Figure 1: Flow chart of the strategy used to identify and analyze shared variants with the various datasets used in this study.

Figure 2: Definition and analysis of the cancer shared dataset. A: Venn diagram highlighting variants shared by the four cancer mutation databases. B: Classification of the 471 variants in the CSD according to ClinVar. C: Boxplot plot analysis of *TP53* variant scores according to SIFT, CADD and REVEL derived from dbSNFP or Envision. The three datasets analyzed correspond to all TP53 variants found in the Sanger dataset (Sanger), in CSD (CSD_IN) or in the Sanger dataset without CSD variants (CSD_OUT). Box and whisker plots show the upper and lower quartiles and range (box), median value (horizontal line inside the box), and 1-99 percentile (whisker line). For SIFT, CADD and REVEL, higher scores are associated with deleterious variants, whereas the opposite applies to ENVISION scores. D: Boxplot plot showing TP53 variant loss of activity in the various datasets.

TP53 mutants are classified into 6 categories according to their frequencies in the database, only using data derived from Sanger studies (orange plot). CSD, gnomAD and gnomAD_CSD data are shown in blue, green and yellow respectively. P values listed above each bar refer to the comparison with the 500+ category. The red line indicates the threshold for TP53 loss of activity i.e. less than 20% compared to wt TP53. For C and D, pairwise Mann–Whitney U test was used to evaluate statistical significance between the various groups. NS, not significant. ***: $p<0.0001$

Figure 3: Identification of deleterious TP53 variants in gnomAD. **A**: The heat maps correspond to the residual transcriptional activity of the 34 TP53 missense variants found in both gnomAD and the CSD. Each column represents a different transcription promoter and each row represents a different *TP53* variant. Activities are displayed from red (lowest) to green (highest). Variant ID using both cDNA (NM_000546.5) and protein (NP_000537.3) nomenclature is shown on the right for each variant. The count of each variant is shown on the right side of the heat map as blue bars for the Sanger database and green bars for gnomAD and ExAC datasets. **B**: same as A with the 8 missense TP53 variants that have been filtered out in version 2.02 of gnomAD, but still included in the ExAc database. (See Material and methods for more information).

Red Stars: last base exon variants known to impair TP53 splicing. Although both variants potentially lead to an amino acid substitution, they should be considered to be TP53 null and pathogenic.

Red arrows: *TP53* variants detected in the ExAC dataset that does not contain TCGA data.

Figure 4: Frequency distribution of *TP53* variants found in the two sets extracted from gnomAD. Frequent variants (right part) are well-known non-pathogenic SNP that have been fully characterized in many studies. Variants in the left part are a mix of extremely rare non-pathogenic and pathogenic variants. Variants in the gray zone (double arrows) will need careful functional evaluation to define their functional status. Green dots, upper part: non pathogenic gnomAD variants; red dots, lower part: pathogenic gnomAD variants.

FIGURE 1

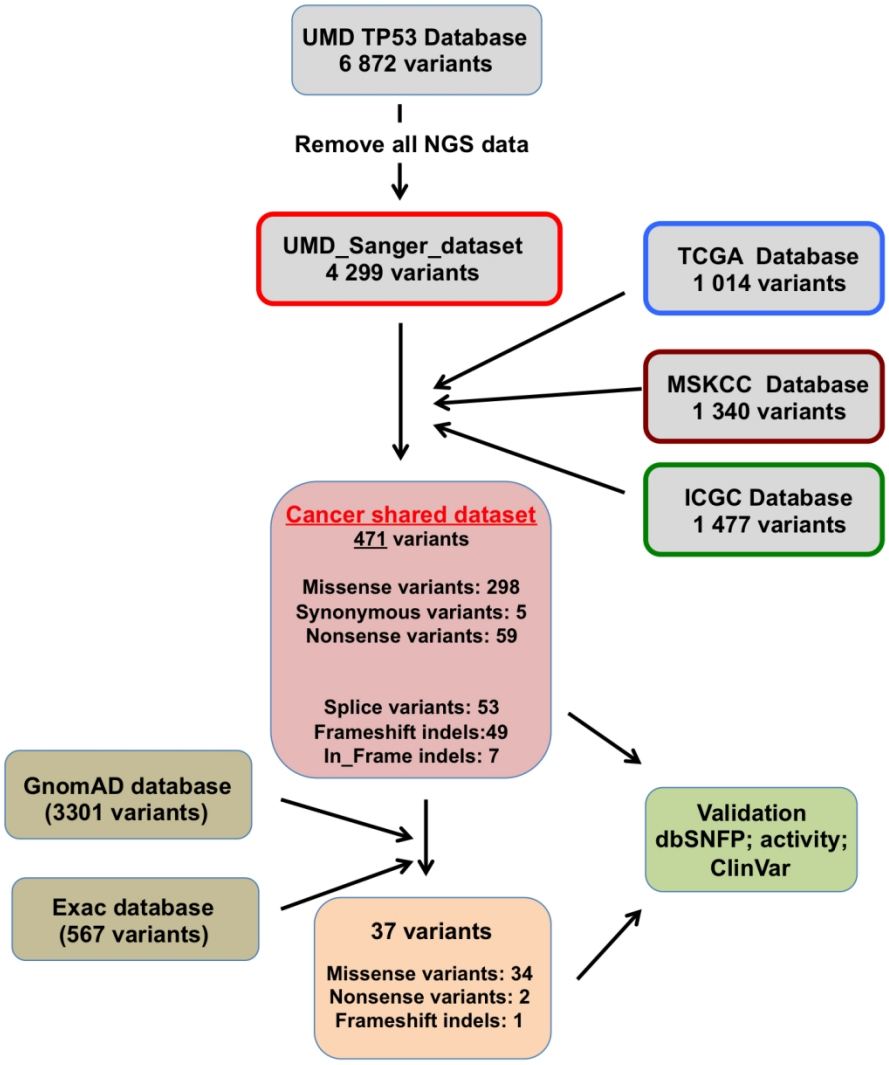

Figure 1

508x677mm (72 x 72 DPI)

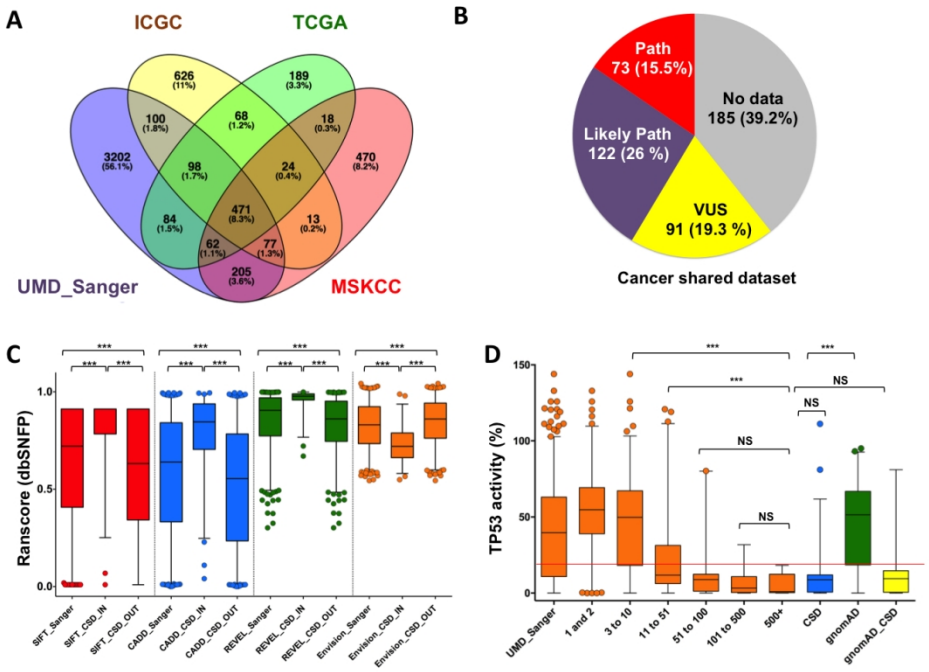

Figure 2

Figure 2

508x381mm (72 x 72 DPI)

	Pathogenic	Likely pathogenic	Uncertain*	Benign	Other**	Total
APC	9	7	594	324	18	949
PTEN	9	1	92	87	0	189
BRCA1	96	1	470	325	1	893
MLH1	11	7	277	163	3	461
STK11	2	2	126	120	0	250
BRCA2	199	5	1218	664	1	2087

Table 1: Summary of ClinVar data for variants found in the gnomAD data set. * ClinVar data were labelled as either uncertain or conflicting.** No prediction data in ClinVar

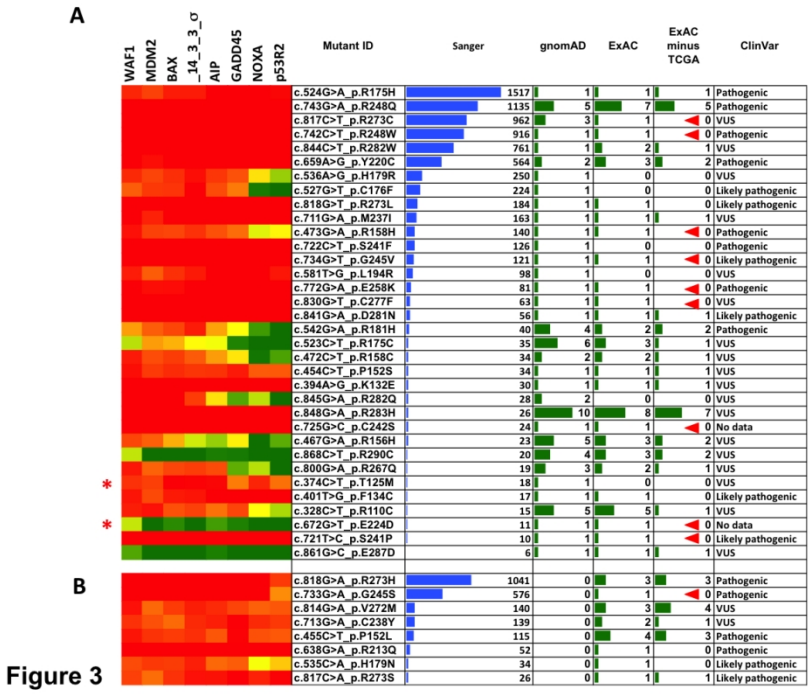

Figure 3

Figure 3

508x381mm (72 x 72 DPI)

Figure S1: Boxplot analysis of TP53 variant scores according to the various predictive criteria derived from dbSNFP as defined in Figure 2 The three datasets analyzed correspond to all TP53 variants found in the Sanger dataset (Sanger), in CSD (CSD_IN) or in the Sanger dataset without CSD variants (CSD_OUT). Box and whisker plots show the upper and lower quartiles and range (box), median value (horizontal line inside the box), and 1-99 percentile (whisker line). Higher scores are associated with deleterious variants.

Figure S2: Heat map showing the residual activity of TP53 missense variants selected (A) or unselected (B) in the cancer shared dataset. Each column represents a different transcription promoter and each row represents a different TP53 variant (see Figure 2 for more information). Activities are displayed from red (lowest) to green (highest). Variants have been ranked according to their frequency in the UMD_TP53 database.

Figure S3: ClinVar classification of TP53 variants. A: pie chart of gnomAD data as defined by ClinVar for all TP53 variants. B: same as A excluding variants with no data. C: classification of gnomAD TP53 variants exclusively found in the CSD.

Figure S4: Boxplot analysis of TP53 variant scores according to the various predictive criteria derived from dbSNFP as defined in material and methods. The three datasets analyzed correspond to TP53 variants found in the entire gnomAD dataset (gnomAD) or in the subset of gnomAD included (gnomAD_CSD_IN) or excluded (CSD_OUT) from the CSD. Box and whisker plots show the upper and lower quartiles and range (box), median value (horizontal line inside the box), and 1-99 percentile (whisker line). Higher scores are associated with deleterious variants.

Figure S5: Heat map showing the residual activity of TP53 variants from the gnomAD dataset that were not detected in the CSD. Each column represents a different transcription promoter and each row represents a different TP53 variant (see material and methods for more information). Activities are displayed from red (lowest) to green (highest). Variants have been ranked according to their frequency in the UMD_TP53 database. Variant ID using both cDNA (NM_000546.5) and protein (NP_000537.3) nomenclature is shown for each variant. The count of each variant is shown on the right side of the heat map as blue and green bars for the Sanger and gnomAD datasets, respectively.

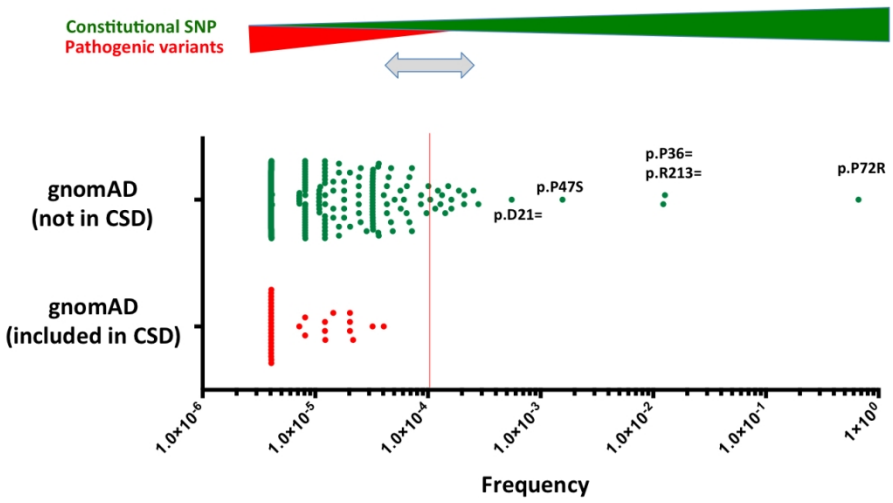

Figure 4

Figure 4

508x381mm (72 x 72 DPI)

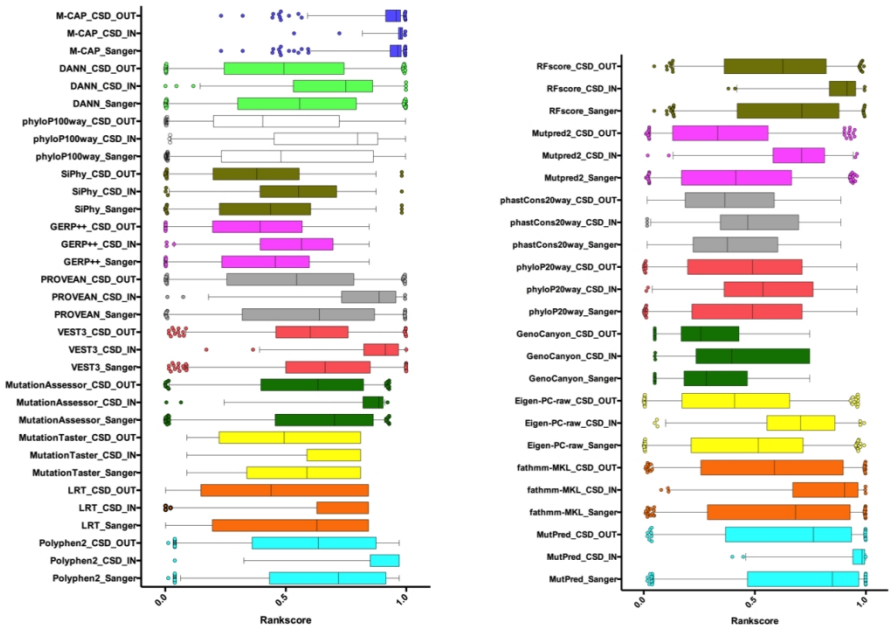

Figure S1

508x381mm (72 x 72 DPI)

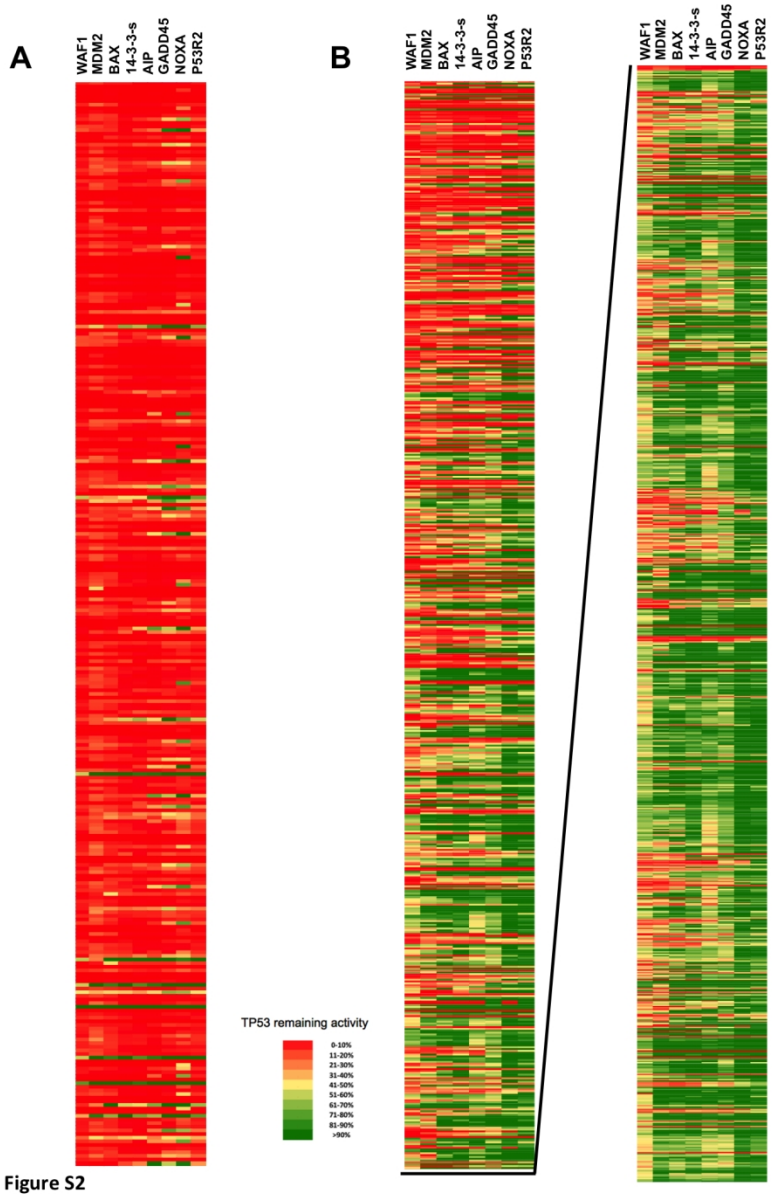

Figure S2

508x733mm (72 x 72 DPI)

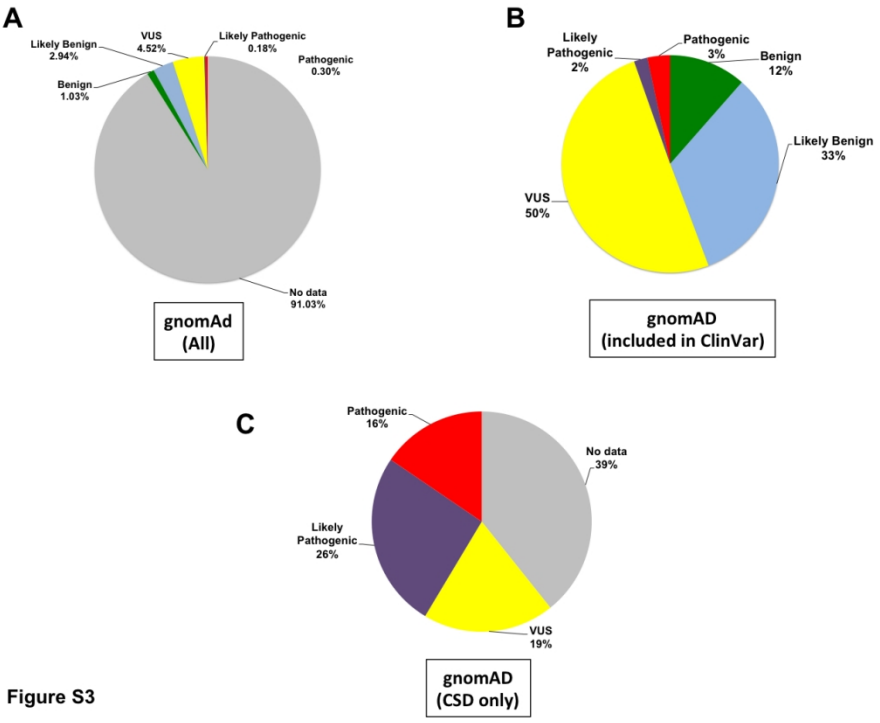

Figure S3

508x381mm (72 x 72 DPI)

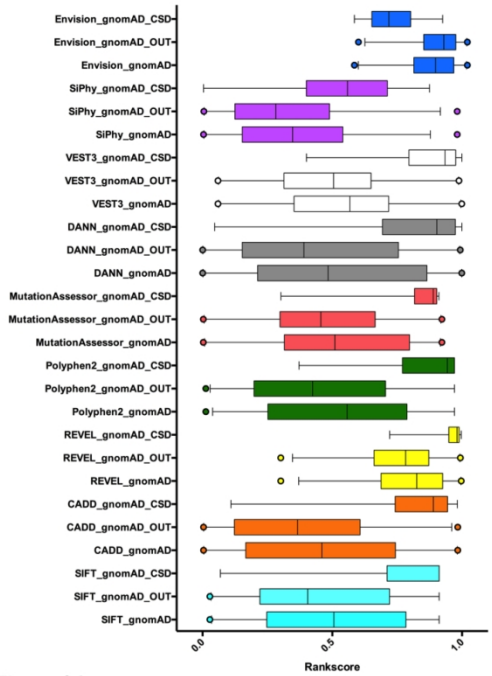

Figure S4

508x381mm (72 x 72 DPI)

Supplementary Table S1

Number of TP53 variants in ExAC and gnomAD datasets according to Ensembl coordinates.

	ExAC	ExAC minus TCGA	gnomAD
Ensembl coordinates* 7,661,779-7,687,550	567	455	1053
RefSeq coordinates* 7,692,550- 7,659,779	567	455	3301

* GRCh38